Verfahren und System zum Erfassen von Daten aus maschinell lesbaren Dokumenten

Die Erfindung betrifft ein Verfahren und ein System zum Erfassen von Daten aus maschinell lesbaren Dokumenten, wobei die Daten einer Datenbank zugeordnet werden, in dem einzelne Daten möglichst automatisch dem Dokument extrahiert werden und in entsprechende Datenbankfelder eingetragen werden, wobei das erfindungsgemäße Verfahren und System das Erfassen von Daten betrifft, falls Daten für ein oder mehrere bestimmte Datenbankfelder eines Dokumentes nicht mit der notwendigen Zuverlässigkeit extrahiert werden konnten.

Verfahren und Systeme zum Erfassen von Daten aus maschinell lesbaren Dokumenten sind bekannt. Üblicherweise weisen die Systeme einen Scanner auf, mit welchem Vorlagen optisch abgetastet werden. Die hierbei erzeugten Dateien sind maschinell lesbare Dokumente und enthalten in der Regel Textelemente. Die Textelemente werden mit Hilfe einer OCR-Einrichtung in codierten Text umgesetzt. Den Dateien werden in der Regel vorbestimmte Formulare bzw. Templates zugeordnet, so dass anhand der Formulare gezielt bestimmte Informationen aus den Text enthaltenden Dateien ermittelt werden können. Diese Informationen werden zum Beispiel in einer Datenbank abgespeichert.

Derartige Verfahren und Systeme werden beispielsweise bei großen Firmen eingesetzt, um Rechnungen zu lesen. Die so extrahierten Daten können automatisch einer betriebswirtschaftlichen Software übermittelt werden.

Ein solches System ist in der US 4,933,979 beschrieben. Dieses System weist einen Scanner zum optischen Abtasten von Formularen auf. Bei diesem System können eine Vielzahl von Formulartypen definiert werden, wobei jeder Formulartyp bzw. Template durch mehrere Parameter, insbesondere geometrisch

definierte Bereiche, in welchen Texte oder Bilder enthalten sein sollen, festgelegt ist. Die Formulartypen können auch durch weitere Eigenschaften, wie zum Beispiel der Schrift, die in den Texten enthalten ist (Alphabet, Zahlen, Symbole, Katakana, Kanji, Handschrift) definiert sein. Nach dem Scannen eines Formulars wird mittels einer Formulartypunterscheidungseinrichtung dem gescannten Formular ein Template zugeordnet. Dementsprechend werden die in dem Textfeld enthaltenen Daten mittels einer OCR-Einrichtung gelesen und extrahiert. Falls kein geeignetes Template vorhanden ist, muss eines erstellt werden.

Aus der WO 98/47098 geht ein weiteres System zum automatischen Erfassen von Daten aus maschinell lesbaren Dokumenten hervor. Hierbei werden mittels eines Scanners Formulare optisch abgetastet. Danach wird automatisch eine Linien-Karte des Formulars erstellt. Hierbei werden zum einen alle Linien erfasst als auch grafische Elemente in eine Linienstruktur umgesetzt. Andere Elemente, wie zum Beispiel Textabschnitte, werden ausgefiltert. Alle vertikalen Linien bilden die Grundlage zur Erstellung eines vertikalen Schlüssels und alle horizontalen Linien bilden die Grundlage zur Erstellung eines horizontalen Schlüssels. Danach wird ermittelt, ob bereits ein Template mit einem korrespondierenden vertikalen und horizontalen Schlüssel vorhanden ist. Falls dies der Fall ist, werden die Daten mit einem entsprechenden Template ausgelesen. Ist dies nicht der Fall, so wird anhand des eingescannten Formulars mittels eines Selbstlern-Modus ein Template erstellt und abgespeichert.

In dem Buch Modern Information Retrieval von Baeza-Yates und Ribeiro-Neto, Eddison-Wessley Verlag, ISBN 0-201-39829-X sind die Grundlagen von Datenbanken und zum schnellen Wiederauffinden von in Datenbanken gespeicherten Informationen erläutert. So ist im Kapitel 8.2 ein Verfahren mit invertierten Dateien, das auch als invertierter Index bezeichnet wird, beschrieben. Bei diesem Verfahren wird aus einem zu untersu-

chenden Text zunächst ein Wörterbuch mit allen im Text enthaltenen Wörtern erstellt. Allen Wörtern des Wörterbuches werden eine oder mehrere Zahlen zugeordnet, die angeben, an welcher Stelle das Wort im Text auftritt. Derartige invertierte Dateien erlauben eine schnellere automatische Analyse eines zu durchsuchenden Textes. Im Kapitel 8.6.1 ist ein String Matching-Verfahren beschrieben, mit welchem zwei Strings verglichen werden und ein zur Ähnlichkeit der Strings indirekt proportionales Kostenmaß berechnet wird. Wenn die beiden Strings identisch sind, ist der Betrag des Kostenmaßes Null. Je stärker sich die Strings unterscheiden, desto größer ist der Betrag des Kostenmaßes. Das Kostenmaß ist somit ein Ausdruck für die Ähnlichkeit der beiden Strings. Dieses und ähnliche Verfahren sind auch unter den Bezeichnungen Approximate String Matching, Levenshtein-Verfahren, Elastic Matching und Viterbi-Algorythmus bekannt. Diese Verfahren gehören zu dem Gebiet der dynamischen Programmierung.

Aus der noch nicht veröffentlichten Patentanmeldung DE 103 42 594.2 geht ein Verfahren und ein System zum Erfassen von Daten aus mehreren maschinell lesbaren Dokumenten hervor, bei dem aus einem zu bearbeitenden Dokument, dem Lesedokument, Daten extrahiert werden, indem sie an Positionen aus dem Lesedokument ausgelesen werden, die durch in einem Vorlagedokument eingetragene Felder bestimmt sind.

Tritt ein Fehler beim Auslesen der Lesedokumente auf, wird das Lesedokument an einem Bildschirm dargestellt und lediglich durch Markieren entsprechender Felder im Lesedokument können die Daten ausgelesen werden. Hierbei werden, falls es erforderlich ist, automatisch weitere Vorlagedokumente anhand der markierten Lesedokumente erstellt bzw. vorhandene Vorlagedokumente entsprechend korrigiert. Dieses System ist derart einfach bedienbar, so dass keine speziellen Computer- oder Softwarekenntnisse notwendig sind.

Der Erfindung liegt die Aufgabe zugrunde, ein Verfahren und ein System zum Erfassen von Daten aus maschinell lesbaren Dokumenten zu schaffen, bei dem die Eingabe der Daten gegenüber dem bekannten Verfahren erheblich vereinfacht wird, falls Daten nicht automatisch extrahiert werden konnten.

Die Aufgabe wird durch ein Verfahren mit den Merkmalen des Anspruchs 1 und durch ein System mit dem Merkmal des Anspruchs 16 gelöst. Vorteilhafte Ausgestaltungen der Erfindung sind in den jeweiligen Unteransprüchen angegeben.

Mit den oben erläuterten Verfahren können Daten aus mehreren maschinell lesbaren Dokumenten erfasst werden, wobei die Daten einer Datenbank zugeordnet werden, indem einzelne Daten möglichst automatisch dem Dokument extrahiert werden und in entsprechende Datenbankfelder eingetragen werden. Falls Daten für ein oder mehrere bestimmte Datenbankfelder einem Dokument nicht mit der notwendigen Zuverlässigkeit extrahiert werden konnten, zum Beispiel weil ein Fehler festgestellt worden ist, der beispielsweise dadurch verursacht sein kann, dass in dem Dokument an der Stelle, wo die Daten gelesen werden sollten, keine Daten oder falsche Daten vorhanden sind, oder dass beim Einlesen dieses Dokumentes mittels eines OCR-Verfahrens ein oder mehrere Zeichen falsch umgesetzt werden, so werden erfindungsgemäß folgende Schritte ausgeführt:

- Darstellen des Dokumentes an einem Bildschirm,
- Anzeigen des Datenbankfeldes, für das die Daten nicht mit der notwendigen Zuverlässigkeit extrahiert werden konnten, am Bildschirm
- Ausführen einer Vorschlags-Routine, mit welcher Stringabschnitte in der Nähe eines von einem Benutzer auf dem Bildschirm bewegbaren Zeiger ausgewählt, markiert und zur Extraktion vorgeschlagen werden.

Das Dokument wird am Bildschirm dargestellt, damit der Benutzer es lesen kann. Zudem wird das Datenbankfeld angezeigt,

für das die Daten nicht mit der notwendigen Zuverlässigkeit extrahiert werden konnten. Hierdurch wird der Benutzer in Kenntnis darüber gesetzt, für welches Datenbankfeld die Daten aus dem am Bildschirm dargestellten Dokument noch zu extrahieren sind.

Durch das Ausführen bzw. Aktivieren der Vorschlags-Routine werden Stringabschnitte in der Nähe eines von dem Benutzer auf dem Bildschirm bewegbaren Zeiger ausgewählt, markiert und zur Extraktion vorgeschlagen. Hierdurch muss der Benutzer lediglich den Zeiger auf dem am Bildschirm dargestellten Dokument in die Nähe eines Stringabschnittes bewegen, der die Daten für das angezeigte Datenbankfeld enthält. Die Daten werden dann automatisch ausgewählt, markiert und zur Extraktion vorgeschlagen. Der Benutzer kann dann lediglich durch Betätigen einer bestimmten Taste den vorgeschlagenen Stringabschnitt in das Datenbankfeld übernehmen.

Durch das automatische Auswählen und Markieren des Stringabschnittes wird der Vorgang der Übernahme der noch fehlenden Daten erheblich vereinfacht und beschleunigt.

Nach einer bevorzugten Ausführungsform der Erfindung werden beim Auswählen des Stringabschnittes Konzept-Informationen berücksichtigt, die dem jeweiligen Datenbankfeld zugeordnet sind.

Die Erfindung wird nachfolgend beispielhaft anhand der Zeichnung näher erläutert. In der Zeichnung zeigen:

Figur 1 ein Verfahren zum Erfassen von Daten aus einem Dokument, die nicht automatisch extrahiert werden konnten,

Figur 2 - 6 jeweils Kopien von Bildschirmdarstellungen zu einzelnen Verfahrensschritten des in Figur 1 gezeigten Verfahrens,

Figur 7 ein Verfahren zum Extrahieren von in Tabellen angeordneten Daten,

Figur 8, 9 jeweils eine Tabelle mit markierten Daten, und

Figur 10 ein System zum Ausführen des erfindungsgemäßen Verfahrens.

Das erfindungsgemäße Verfahren zum Erfassen von Daten aus maschinell lesbaren Dokumenten ist eine Weiterbildung der eingangs erläuterten Verfahren, mit welchen aus Dokumenten maschinell Daten extrahiert und in einer Datenbank gespeichert werden können.

Bei diesen Verfahren können jedoch nicht immer alle Datenbankfelder der Datenbank zuverlässig mit aus den Dokumenten extrahierten Daten gefüllt werden. Liegt zum Beispiel beim Extrahieren der Daten ein Fehler vor, so wird das automatische Verfahren unterbrochen und unter Mitwirkung eines Benutzers werden die Daten aus dem Dokument manuell in Datenbankfelder übertragen. Ein solcher Fehler kann dadurch verursacht sein, dass in dem zu bearbeitenden Dokument kein geeigneter Stringabschnitt gefunden wird, aus dem die Daten gelesen werden können oder der Stringabschnitt fehlerhafte Daten enthält, die beispielsweise beim Umsetzen in codierten Text des Dokumentes mittels eines OCR-Verfahrens entstanden sind.

Das erfindungsgemäße Verfahren beginnt somit dann, wenn Daten nicht zuverlässig extrahiert werden können. Der Ausdruck „nicht zuverlässig extrahierbar" umfasst sowohl grundsätzliche Fehler beim Lesen von Daten, die ein Lesen der Daten nicht möglich machen als auch gelesene Daten, die zum Beispiel unter Berücksichtigung von Kontext-Informationen auf das Datenbankfeld abgebildet werden, wobei die Güte der Abbildung ermittelt wird. Derartige Abbildungsverfahren sind zum Beispiel das eingangs erläuterte String Matching-

Verfahren. Ist die hierbei erzielte Abbildungsgüte zu gering, so werden die automatisch eingelesenen Daten als nicht ausreichend zuverlässig bewertet und verworfen.

Nachfolgend wird das erfindungsgemäße Verfahren anhand des in Figur 1 dargestellten Flussdiagramms erläutert. In dem Flussdiagramm sind alle Schritte, die automatisch ausgeführt werden, mit einem „a" im Kreis und alle vom Benutzer manuell zu tätigenden Schritte mit einem „m" im Kreis gekennzeichnet.

Es beginnt mit dem Schritt S1.

Nachdem zumindest Daten für ein Datenbankfeld nicht mit der notwendigen Zuverlässigkeit extrahiert werden konnten, wird das entsprechende Dokument 1 an einem Bildschirm 2 dargestellt und das Datenbankfeld 3 angezeigt (Schritt S2). Figur 2 zeigt eine Bildschirmdarstellung unmittelbar nach dem Feststellen, dass Daten nicht mit der notwendigen Zuverlässigkeit extrahiert werden konnten, wobei in einem Fenster 4/1 auf der rechten Seite der Bildschirmdarstellung das Dokument 1 dargestellt ist. Auf der linken Seite sind zwei Fenster 4/2 und 4/3 angeordnet. Das Fenster 4/2 enthält eine Übersicht der zu bearbeitenden Dokumente und im Fenster 4/3 sind die einzelnen Datenbankfelder angegeben, in welche Daten gespeichert werden, die aus dem Dokument 1 zu lesen sind.

In dem dargestellten Beispiel konnte keines der Datenbankfelder mit Daten gefüllt werden, weshalb die einzelnen Datenbankfelder 3 mit dem Zusatz [empty] versehen sind. Es ist jedoch auch möglich, dass nur in wenigen Datenbankfeldern oder lediglich in einem einzigen Datenbankfeld Daten fehlen.

In Figur 2 ist das Datenbankfeld „InvoiceNumber" (= Rechnungsnummer) im Vergleich zu den anderen Datenbankfeldern 3 dunkler markiert, was dem Benutzer anzeigt, dass für dieses Datenbankfeld 3 Daten aus dem Dokument 1 zu extrahieren sind. Zusätzlich ist im oberen Bereich des Fensters 4/1 in großer

Schrift der Begriff „InvoiceNumber" aufgeführt, das dem Benutzer zusätzlich anzeigt, für welches Datenbankfeld Daten zu extrahieren sind.

Der Benutzer kann nun im Fenster 4/1 einen Zeiger 5 positionieren, den er vorzugsweise derart anordnet, dass er sich möglichst nahe an dem Stringabschnitt befindet, von dem der Benutzer annimmt, dass dessen Inhalt in dem entsprechenden Datenbankfeld abzuspeichern ist. In dem in Figur 2 gezeigten Beispiel sind Daten für das Datenbankfeld „Rechnungsnummer" zu extrahieren, weshalb der Zeiger 5 in der Nähe der Rechnungsnummer „4361" positioniert wird (Schritt S3).

Der Zeiger 5 kann hierbei mittels einer Maus 6 oder durch Eingabe an einer Tastatur 7 im Fenster 4/1 bewegt werden.

Nach dem Positionieren des Zeigers 5 beginnt eine Vorschlags-Routine, die mehrere Verfahrensschritte umfasst. Diese Vorschlags-Routine kann einerseits dadurch ausgelöst werden, dass der Zeiger 5 ein vorbestimmtes Zeitintervall nicht bewegt wird, wodurch dann die Vorschlags-Routine automatisch ausgeführt wird, oder dadurch, dass eine bestimmte Taste einer Maus oder der Tastatur bestätigt wird.

Im Schritt S4 wird zunächst geprüft, ob in der näheren Umgebung des Zeigers ein Stringabschnitt mit einem für das Datenbankfeld 3 geeigneten Konzept vorhanden ist, sofern dem entsprechenden Typ des Datenbankfeldes vorab Konzeptinformationen zugeordnet wurden. Diese Konzeptinformationen umfassen die Syntax und/oder die Semantik des Datenbankfeldes. Informationen zur Syntax sind zum Beispiel die Anzahl von Ziffern und/oder Buchstaben und/oder vorbestimmte Formate des zu lesenden Stringabschnittes. So weisen Datumsfelder, Betragsfelder und Adressfelder in der Regel bestimmte Formate auf. Informationen zur Semantik umfassen vorbestimmte Begriffe, die in das entsprechende Datenbankfeld eingefügt werden können. Dies ist zum Beispiel bei Währungsangaben zweckmäßig, oder

wenn die Artikelbezeichnung eines bestimmten Lieferanten eingelesen werden sollen, der eine begrenzte Anzahl von Artikeln liefern kann. Die entsprechenden Artikelbezeichnungen sind dann in einem Lexikon abgelegt und können dann eindeutig erkannt werden.

Bei dem in Figur 2 dargestellten Ausführungsbeispiel befinden sich in der Nähe des Zeigers 5 die zwei Stringabschnitte „4361" und „02.08.2002". Der letzte Stringabschnitt besitzt die Syntax eines Datums, weshalb er zum Extrahieren der Rechnungsnummer verworfen wird. Der Stringabschnitt „4361" entspricht der Syntax einer Rechnungsnummer. Somit wird im Schritt S4 entschieden, dass ein Stringabschnitt mit einem geeigneten Konzept vorliegt, weshalb der Verfahrensablauf auf den Schritt S5 übergeht. Im Schritt S5 wird der Stringabschnitt „4361" markiert (Figur 3). Die Markierung erfolgt im vorliegenden Ausführungsbeispiel durch eine farbliche Unterlegung des Stringabschnittes und durch Zeichnen eines Rahmens 8.

Sollte im Schritt S4 kein geeignetes Konzept ermittelt werden, so geht der Verfahrensablauf auf den Schritt S6 über. Im Schritt S6 wird das zum Zeiger 5 nächstliegend angeordnete Einzelzeichen ermittelt, das im vorliegenden Ausführungsbeispiel gemäß Fig. 2 - 4 die „1" ist. Danach werden die Grenzen des dieses Zeichen enthaltenden Stringabschnittes nach allgemeinen Regeln ermittelt. Diese Grenzen können zum Beispiel durch Leerzeichen bzw. Leerräume im Dokument 1 oder durch bestimmte Satzzeichen oder sonstige Markierungen im Dokument 1 vorgegeben sein. Werden entsprechende Begrenzungsmarkierungen erkannt, so wird der dazwischen liegende Stringabschnitt ausgewählt und markiert. Bei dem in Figur 2 und 3 gezeigten Ausführungsbeispiel befinden sich seitlich des Stringabschnittes „4361" jeweils Leerräume, durch die auch nach den allgemeinen Regeln eine eindeutige Auswahl der Markierung des Stringabschnittes möglich ist.

Unabhängig davon, ob der Stringabschnitt gemäß dem Schritt S5 oder gemäß dem Schritt S6 ausgewählt oder markiert worden ist, geht der Verfahrensablauf auf den Verfahrensschritt S7 über, mit der der Stringabschnitt in einem zusätzlichen Rahmen 9 als codierter Text und in einem weiteren Rahmen 10 vergrößert dargestellt wird (Fig. 3, 4). Bei dem vorliegenden Ausführungsbeispiel liegt das Dokument 1 als grafische Datei, zum Beispiel im pdf, tif, gif, jpg-Format vor. Üblicherweise wird in dem bereits vorausgegangenen Verfahrensabschnitt das Dokument einer OCR-Routine unterzogen und in codierten Text umgesetzt. Der codierte Text ist hierbei auch auf Konzepte untersucht worden und die entsprechenden Informationen sind abgespeichert worden. Diesem codierten Text wird der zu dem Stringabschnitt korrespondierende Abschnitt entnommen und in dem Rahmen 9 dargestellt. Hierdurch erkennt der Benutzer, ob der Stringabschnitt korrekt in codierten Text umgesetzt worden ist.

Im Rahmen 10 wird der Stringabschnitt im Grafikformat in vergrößerter Darstellung dargestellt, wodurch der Benutzer auch Details im Stringabschnitt erkennen kann.

Mit dem Schritt S7 ist die Vorschlags-Routine abgeschlossen.

Im Schritt S8 beurteilt der Benutzer, ob der ausgewählte und markierte Stringabschnitt grundsätzlich zur Übernahme in das Datenbankfeld geeignet ist. Ist dies nicht der Fall, so wird der Zeiger 5 erneut positioniert (S3) und die Vorschlags-Routine (S4 - S7) wiederholt ausgeführt. Ist die Auswahl des Stringabschnitts hingegen grundsätzlich geeignet, so beurteilt der Benutzer, ob auch der markierte Bereich korrekt ist (Schritt S9). Ist dies nicht der Fall, so kann der Benutzer die Markierung des Stringabschnittes manuell bearbeiten und/oder den codierten Text im Rahmen 9 editieren (Schritt S10). Mit dem Editieren des codierten Textes können Fehler, die durch eine nicht korrekte OCR-Umsetzung entstanden sind, behoben werden. Bei diesen Korrekturen (Bereich anpassen, E-

ditieren) werden automatisch der markierte Bereich und die Inhalte der Rahmen 9 und 10 angepasst.

Ist der markierte Bereich korrekt bzw. vom Benutzer entsprechend überarbeitet worden, so geht der Verfahrensablauf auf den Schritt S11 über, mit dem die in dem ausgewählten Stringabschnitt enthaltenen Daten in das korrespondierende Datenbankfeld übertragen werden (Figur 4). Diese Übertragung der Daten wird durch eine Betätigung einer vorbestimmten Taste an der Maus oder der Tastatur vom Benutzer ausgelöst. Danach ist das Verfahren zum Extrahieren von Daten für ein Datenbankfeld beendet (S12). Sind Daten für weitere Datenbankfelder zu lesen, so beginnt das Verfahren erneut mit dem Schritt S1. In Figur 5 ist das nächste zu lesende Datenbankfeld „Invoice Date" (= Rechnungsdatum) angezeigt.

Mit dem erfindungsgemäßen Verfahren wird die Tätigkeit eines Benutzers beim manuellen Übertragen von Daten aus einem Dokument in ein Datenbankfeld lediglich auf das Positionieren des Zeigers, der Kontrolle der automatisch vorgeschlagenen Auswahl und der eventuellen Korrektur des Bereiches und das Betätigen einer Taste zum Übertragen der Daten beschränkt. Die Auswahl und die Markierung des Bereichs des auszuwählenden Stringabschnittes wird vom erfindungsgemäßen Verfahren selbsttätig ausgeführt.

Die Figuren 2 bis 5 zeigen die Übernahme von Daten in ein einzelnes Datenbankfeld. Durch die Berücksichtigung von Konzept-Informationen ist es jedoch auch möglich, mit einem einzigen Stringabschnitt Daten für mehrere Datenbankfelder zu extrahieren. Figur 6 zeigt ein entsprechendes Ausführungsbeispiel, in dem die vollständige Adresse als ein Stringabschnitt markiert und gelesen wird, wobei die Adresse selbsttätig in die einzelnen Datenbankfelder, Name, Firma, Straße, Postleitzahl und Stadt segmentiert wird.

Nachfolgend wird eine weitere Ausgestaltung des oben beschriebenen Verfahrens anhand des Flussdiagramms aus Figur 7 und der Bildschirmdarstellungen gemäß Figur 8 und 9 erläutert, mit dem Daten aus Tabellen extrahiert werden können.

Dieses Verfahren beginnt mit dem Schritt S15.

Im Schritt S16 werden die Werte der Tabelle in der ersten Tabellenzeile gemäß obigen Verfahren durch Positionieren des Zeigers, automatisches Auswählen und Markieren des Stringabschnittes und Übernehmen der Daten in korrespondierende Datenbankfelder extrahiert. Figur 8 zeigt eine Tabelle, in der die Stringabschnitte der ersten Tabellenzeile markiert sind, die in die entsprechenden Datenbankfelder übernommen worden sind. Diese Datenbankfelder besitzen die Struktur einer Tabelle, zum Beispiel sind sie als zwei-dimensionales Datenfeld angelegt, so dass beim Extrahieren der Daten in diese Datenbankfelder das Verfahren anhand des Datenbankfeldes selbsttätig erkennt, dass Daten aus einer Tabelle ausgelesen werden.

Eine Tabellenzeile kann sich auch über mehrere Seiten erstrecken, wenn die Tabelle sich entsprechend über mehrere Seiten erstreckt. Sind die Daten der ersten Tabellenzeile vollständig extrahiert worden, kann der Benutzer durch eine vorbestimmte Eingabe das automatische Extrahieren der weiteren Tabelleneinträge initiieren. Wird diese Eingabe vom Benutzer getätigt, so werden im Schritt S17 zunächst eine Liste mit allen Stringabschnitten erstellt, die unterhalb der ersten Tabellenzeile angeordnet sind.

In Schritt S18 wird mittels einer Kostenfunktion ein Kostenwert zwischen Sequenzen von Stringabschnitten der Liste und der Sequenz der Stringabschnitte der ersten Tabellenzeile, auf deren Grundlage Daten in die Datenbankfelder im Schritt S16 extrahiert worden sind, ermittelt. Bei dieser Kostenfunktion werden den Sequenzen der Stringabschnitte der Liste geringe Kosten zugewiesen, deren Stringabschnitte bezüglich ih-

rer horizontalen Position und ihrer Breite mit den korrespondierenden Stringabschnitten der ersten Tabellenzeile übereinstimmen oder zumindest sehr ähnlich sind. Dieser Kostenwert ist somit indirekt proportional zur Ähnlichkeit zwischen den in der Liste aufgeführten Sequenzen von Stringabschnitten und den in der ersten Tabellenzeile enthaltenen Sequenz von Stringabschnitten.

Die hierbei verwendete Kostenfunktion entspricht der aus dem Kapitel 8.6.1 String Matching Allowing Errors in Modern Information Retrieval (ISBN 0-201-39829-X) beschriebenen Kostenfunktion, mit welcher jeweils ein Einzelkostenwert zwischen einem Stringabschnitt der ersten Tabellenzeile und einem Stringabschnitt der weiteren Tabellenzeilen ermittelt wird. Da jede Sequenz mehrere Stringabschnitte umfasst, wird mittels des Viterbi-Algorithmus für die einzelnen Sequenzen von Stringabschnitten jeweils ein Gesamtkostenwert bzw. Gesamtähnlichkeitswert durch Summieren der Einzelkostenwerte berechnet.

Anhand dieser Kostenwerte bzw. Ähnlichkeitswerte werden die Sequenzen von Stringabschnitten als Tabellenzeilen bestimmt, deren Ähnlichkeitswert unter einem vorbestimmten Schwellwert liegt (S19). Hierdurch sind alle Tabellenzeilen und damit Tabelleneinträge der Tabelle bestimmt. Sie werden im Schritt S20 markiert (Figur 9) und im Schritt S21 extrahiert, d.h., automatisch ausgelesen, gegebenenfalls in codierten Text umgesetzt, und in den entsprechenden Datenbankfeldern gespeichert.

Mit dem Schritt S22 ist dieses Verfahren beendet.

Zweckmäßigerweise ist es möglich, die Tabelleneinträge nachzubearbeiten, d.h., die markierten Bereiche zu verändern (verschieben, vergrößern, verkleinern) bzw. einzelne Zeilen zu entfernen bzw. hinzuzufügen. Bei einer Nachbearbeitung

werden die Einträge in den Datenbankfeldern automatisch entsprechend angepasst.

Zusätzlich kann beim Auslesen der Daten und Eintragen in die Datenbankfelder eine zusätzliche Kontrolle durch eine Abbildung mittels dem String Matching-Verfahren erfolgen, mit welcher bestimmt wird, wie gut die Einträge mit dem durch die einzelnen Datenbankfelder vorgegebenen Konzept übereinstimmen.

Weiterhin kann das erfindungsgemäße Verfahren mit dem in der deutschen Patentanmeldung DE 103 42 594.2 beschriebenen Verfahren zum Erfassen von Daten aus mehreren maschinell lesbaren Dokumenten kombiniert werden, weshalb auf die Patentanmeldung vollinhaltlich Bezug genommen wird und sie durch Bezugnahme in die vorliegende Patentanmeldung inkorporiert wird.

Bei diesem Verfahren zum automatischen Erfassen von Daten aus mehreren maschinell lesbaren Dokumenten werden Vorlagedokumente mit einem Lesedokument verglichen und deren Ähnlichkeit bewertet. Das hierbei angewandte Verfahren kann auch zum Auslesen einer Tabelle angewendet werden, wobei die Sequenz der ausgewählten Stringabschnitte der ersten Tabellenzeile dem Vorlagedokument entsprechen und die Kombinationen von Stringabschnitten der weiteren Tabellenzeilen den Lesedokumenten entsprechen.

Bei dem oben beschriebenen erfindungsgemäßen Verfahren zum Extrahieren von Daten aus Tabellen muss ein Benutzer lediglich den Zeiger zu den Tabelleneinträgen in der ersten Tabellenzeile bewegen und die Übernahme der dann automatisch ausgewählten und markierten Stringabschnitte als Daten für das entsprechende Datenbankfeld bestätigen. Nachdem er dies für alle Tabelleneinträge der ersten Tabellenzeile ausgeführt hat, muss er lediglich durch eine Eingabe das vollständige Auslesen der weiteren Tabelleneinträge initiieren. Das Ver-

fahren ermittelt dann selbstständig die weiteren Tabelleneinträge, markiert sie und extrahiert die Daten in die Datenbank.

Dies beschleunigt erheblich das Auslesen von Daten aus der Tabelle in eine Datenbank. Der Verfahrensabschnitt S17 bis S21 stellt daher eine eigenständige Erfindung dar, die jedoch bevorzugt in Kombination mit dem in Figur 1 dargestellten Verfahren, auf der sich der Schritt S16 bezieht, verwendet wird.

Figur 10 zeigt schematisch ein System zum Ausführen des erfindungsgemäßen Verfahrens. Dieses System 11 weist einen Computer 12 mit einer Speichereinrichtung 13, mit einer zentralen Prozessoreinrichtung (CPU) 14 und eine Interface-Einrichtung 15 auf. Am Computer 12 sind ein Scanner 16, ein Bildschirm 2 und eine Eingabeeinrichtung 17 angeschlossen. Die Eingabeeinrichtung 17 umfasst eine Tastatur 7 und/oder eine Maus 6.

In der Speichereinrichtung 13 ist ein Softwareprodukt zum Ausführen des erfindungsgemäßen Verfahrens gespeichert, das an der CPU 14 ausgeführt wird. Mit dem Scanner 16 werden Dokumente erfasst und in eine elektronische Datei umgewandelt. Diese elektronischen Dateien werden vom Computer 12 eingelesen und eventuell mittels einer OCR-Routine und/oder eines Verfahrens zum Erkennen spezieller Syntax oder Semantik in der Datei vorverarbeitet. Danach werden die in den Dateien enthaltenen Dokumente entsprechen dem oben beschriebenen Verfahren mit dem System 11 bearbeitet. An der Eingabeeinrichtung 17 können die entsprechenden Eingaben vorgenommen werden, wobei diese sich lediglich auf Bewegen des Zeigers 5 und ein paar wenige Tastatureingaben beschränken. Gegebenenfalls werden die markierten Felder mittels Tastatur oder Mauseingabe verschoben oder durch Vergrößerung bzw. Verkleinerung angepasst oder der codierte Text editiert.

Die Erfindung ist oben anhand eines Ausführungsbeispiels erläutert worden. Im Rahmen der Erfindung sind hiervon Abwandlungen möglich. So kann zum Beispiel anstelle des Rahmens 8 lediglich der Rahmen 10 vorgesehen werden, in dem der ausgewählte Stringabschnitt vergrößert dargestellt wird. Dieser Rahmen 10 stellt auch eine Markierung des Stringabschnittes dar.

Bei dem oben erläuterten Ausführungsbeispiel werden die Dokumente eingescannt und liegen dann in einem Graphikformat vor. Das erfindungsgemäße Verfahren kann jedoch auch zum Lesen von Informationen aus Dokumenten angewandt werden, die bereits in codiertem Text vorliegen, wie zum Bespiel E-Mails. Bei einer solchen Anwendung ist es selbstverständlich nicht notwendig, dass die Dokumente mittels einer OCR-Routine in codierten Text umgesetzt werden.

Die Erfindung kann folgendermaßen kurz zusammengefasst werden:

Die Erfindung betrifft ein Verfahren zum Erfassen von Daten aus maschinell lesbaren Dokumenten, wobei die Daten einer Datenbank zugeordnet werden.

Mit der Erfindung werden Stringabschnitte, die sich in der Nähe eines vom Benutzer bewegbaren Zeigers befinden, automatisch ausgewählt, markiert und deren Inhalt zur Übernahme in eine Datenbank vorgeschlagen.

Nach einer Weiterbildung des erfindungsgemäßen Verfahrens kann der Inhalt einer Tabelle vollautomatisch ausgelesen werden, wenn die Tabelleneinträge in einer ersten Tabellenzeile einmal gemäß obigen Verfahren ausgelesen worden sind.

Es wurden Ausführungsbeispiele der Erfindung beschrieben. Dabei ist klar, dass der Fachmann jederzeit Abwandlungen und Weiterbildungen angeben kann, die das erfindungsgemäße Kon-

zept benutzten. Weiterhin kann die Erfindung sowohl mittels elektronischen Komponenten (Hardware) als auch durch Computerprogrammelemente (Software oder Softwaremodule) realisiert werden. Die Erfindung wird dabei insbesondere aus einer Kombination von elektronischen Hardware-Elementen und Softwareelementen realisiert. Dementsprechend erfaßt die Erfindung auch Computerprogrammprodukte wie z.B. elektronische Datenträger (CD, DVD, Diskette, Bandlaufwerke) bzw. Komponenten, die über Computernetzwerke (Internet) verbreitet werden und/oder auf Computern und insbesondere im Zwischenspeichern geladen, bereit gehalten und/oder zum Ablauf gebracht werden.

Bezugszeichenliste

1 Dokument
2 Bildschirm
3 Datenbankfeld
4 Fenster
5 Zeiger
6 Maus
7 Tastatur
8 Rahmen
9 Rahmen
10 Rahmen
11 System
12 Computer
13 Speichereinrichtung
14 CPU
15 Interface-Einrichtung
16 Scanner
17 Eingabeeinrichtung

Ansprüche

1. Verfahren zum Erfassen von Daten aus maschinell lesbaren Dokumenten, wobei die Daten einer Datenbank zugeordnet werden, indem einzelne Daten möglichst automatisch dem Dokument extrahiert werden und in entsprechende Datenbankfelder eingetragen werden, und falls Daten für ein oder mehrere bestimmte Datenbankfelder einem Dokument nicht mit der notwendigen Zuverlässigkeit extrahiert werden konnten, werden folgende Schritte ausgeführt:
   - Darstellen des Dokumentes an einem Bildschirm,
   - Anzeigen des Datenbankfeldes, für das die Daten nicht mit der notwendigen Zuverlässigkeit extrahiert werden konnten, am Bildschirm,
   - Ausführen einer Vorschlags-Routine, mit welcher Stringabschnitte in der Nähe eines von einem Benutzer auf dem Bildschirm bewegbaren Zeiger ausgewählt, markiert und zur Extraktion vorgeschlagen werden.

2. Verfahren nach Anspruch 1,
   d a d u r c h g e k e n n z e i c h n e t,
   dass der Stringabschnitt nach Maßgabe von dem Datenbankfeld zugeordneten Konzept-Informationen ausgewählt, markiert und zur Extraktion vorgeschlagen wird.

3. Verfahren nach Anspruch 2,
   d a d u r c h g e k e n n z e i c h n e t,
   dass die Konzept-Informationen die Syntax und/oder die Semantik des Datenbankfeldes beschreiben, so dass von der Vorschlags-Routine ein zu markierender Stringabschnitt entsprechend der Syntax bzw. der Semantik des jeweiligen Datenbankfeldes ausgewählt und markiert wird.

4. Verfahren nach Anspruch 3,
d a d u r c h g e k e n n z e i c h n e t,
dass die Informationen zur Syntax die Anzahl von Ziffern und/oder Buchstaben und/oder vorbestimmte Formate des zu lesenden Stringabschnitts beschreiben.

5. Verfahren nach Anspruch 3 oder 4,
d a d u r c h g e k e n n z e i c h n e t,
dass die Informationen zur Semantik vorbestimmte Begriffe, bspw. mit einem Lexikon, beschreiben.

6. Verfahren nach einem der Ansprüche 1 bis 5,
d a d u r c h g e k e n n z e i c h n e t,
dass ein Stringabschnitt ausgewählt wird, der zwischen zwei Begrenzungszeichen angeordnet ist.

7. Verfahren nach Anspruch 6,,
d a d u r c h g e k e n n z e i c h n e t,
dass Begrenzungszeichen Leerzeichen und/oder Satzzeichen umfassen.

8. Verfahren nach einem der Ansprüche 1 bis 7,
d a d u r c h g e k e n n z e i c h n e t,
dass der Text von Dokumenten in grafischer Darstellung zunächst mit einem OCR-Verfahren in codierten Text umgesetzt wird und die Vorschlags-Routine zusätzlich zum markierten Stringabschnitt in grafischer Darstellung den codierten Text dieses Stringabschnittes darstellt.

9. Verfahren nach einem der Ansprüche 1 bis 7,
d a d u r c h g e k e n n z e i c h n e t,
dass zusätzlich zum markierten Stringabschnitt dieser Stringabschnitt nochmals in vergrößerter Darstellung am Bildschirm angezeigt wird.

10. Verfahren nach einem der Ansprüche 1 bis 9, dadurch gekennzeichnet, dass die Vorschlags-Routine nach dem Markieren eines Stringabschnittes eine Funktion aktiviert, mit welcher durch Betätigen einer oder mehrerer vorbestimmter Tasten der Inhalt des markierten Stringabschnittes in die Datenbank übernommen wird.

11. Verfahren nach einem der Ansprüche 1 bis 10, dadurch gekennzeichnet, dass beim Ausführen der Vorschlags-Routine nach dem Bewegen des Zeigers ein vorbestimmtes Zeitintervall abgewartet wird, innerhalb dessen der Zeiger nicht bewegt werden darf, bevor ein Stringabschnitt ausgewählt wird.

12. Verfahren zum Erfassen von Daten aus maschinell lesbaren Dokumenten, wobei die Daten einer Datenbank zugeordnet werden, insbesondere nach einem der Ansprüche 1 - 11, dadurch gekennzeichnet, dass nach dem Einlesen von Daten aus einer ersten Tabellenzeile in korrespondierende Datenbankfelder automatisch die weiteren Tabelleneinträge durch einen Vergleich von unterhalb der letzten Tabellenzeile angeordneten Stringabschnitten mit den Stringabschnitten der ersten Tabellenzeile, die den eingelesenen Daten entsprechen, ermittelt werden und diese weiteren Tabelleneinträge automatisch extrahiert werden.

13. Verfahren nach Anspruch 12, dadurch gekennzeichnet, dass der Vergleich zwischen den Stringabschnitten mittels eines String Matching-Verfahrens erfolgt.

14. Verfahren nach Anspruch 12 oder 13, dadurch gekennzeichnet, dass die ermittelten Tabelleneinträge markiert werden.

15. Verfahren nach Anspruch 14,
d a d u r c h g e k e n n z e i c h n e t,
dass Funktionen zum Editieren der markierten Tabelleneinträge bereitgestellt werden.

16. System zum Erfassen von Daten aus maschinell lesbaren Dokumenten umfassend einen Computer (12) mit einer Speichereinrichtung (13) und eine CPU (14), wobei in der Speichereinrichtung (13) ein Softwareprodukt zum Ausführen des Verfahrens nach einem der Ansprüche 1 - 15 gespeichert ist.

17. System nach Anspruch 16,
d a d u r c h g e k e n n z e i c h n e t,
dass das System eine Eingabeeinrichtung (17) in Form einer Maus (6) und/oder Tastatur (7) aufweist.

18. System nach Anspruch 16 oder 17,
d a d u r c h g e k e n n z e i c h n e t,
dass das System einen Scanner (16) zum optischen Abtasten von Dokumenten aufweist.

19. Computerprogrammprodukt, das bei seinem Laden und Ausführen auf einem Computer (12) ein Verfahren nach einem der Ansprüche 1 bis 15 bewirkt.

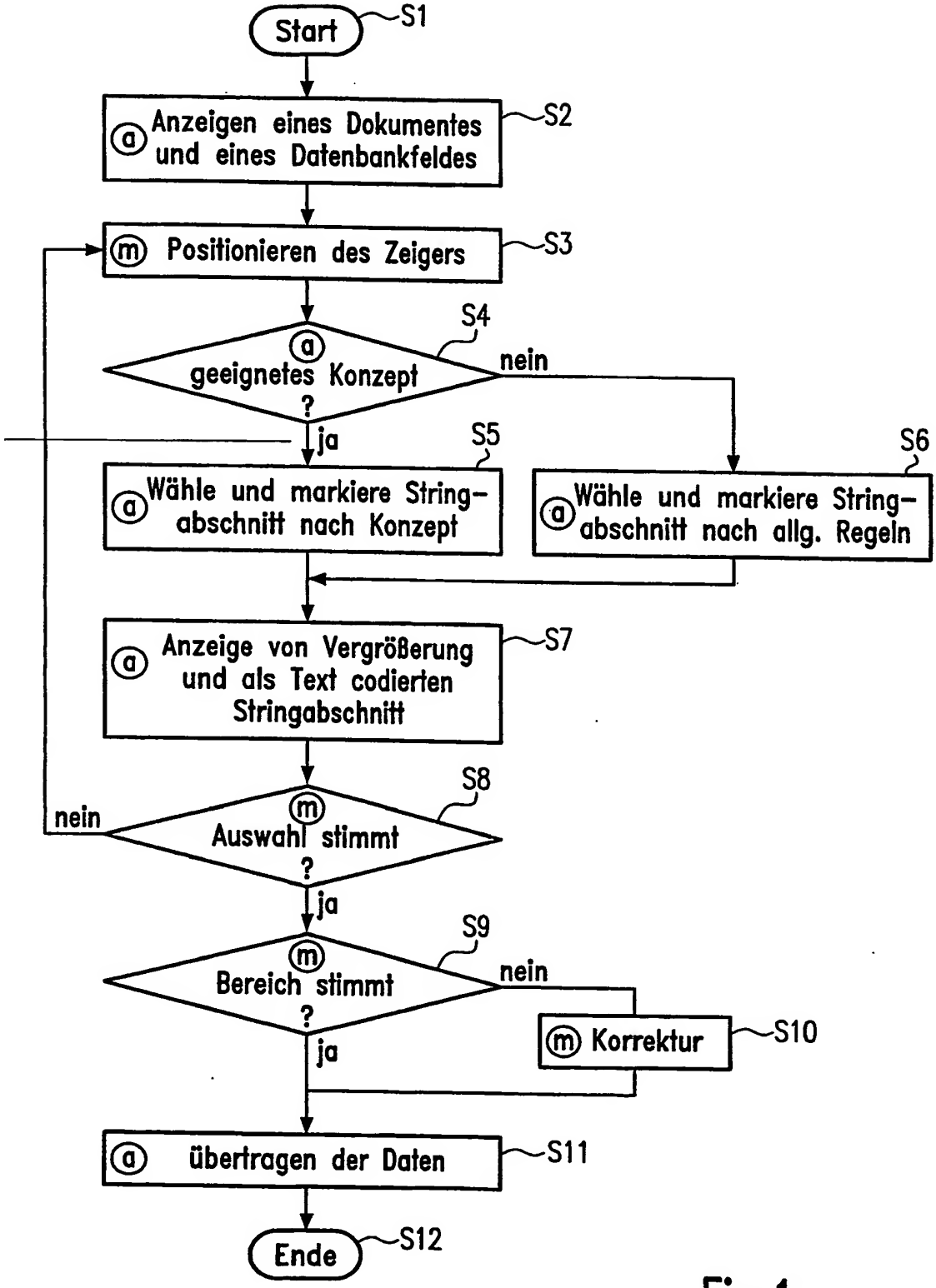

Start
S1
Anzeigen eines Dokumentes und eines Datenbankfeldes
S2
Positionieren des Zeigers
S3
S4
geeignetes Konzept ?
nein
ja
S5
Wähle und markiere String-abschnitt nach Konzept
S6
Wähle und markiere String-abschnitt nach allg. Regeln
Anzeige von Vergrößerung und als Text codierten Stringabschnitt
S7
S8
Auswahl stimmt ?
nein
ja
S9
Bereich stimmt ?
nein
ja
Korrektur
S10
übertragen der Daten
S11
Ende
S12

Fig.1

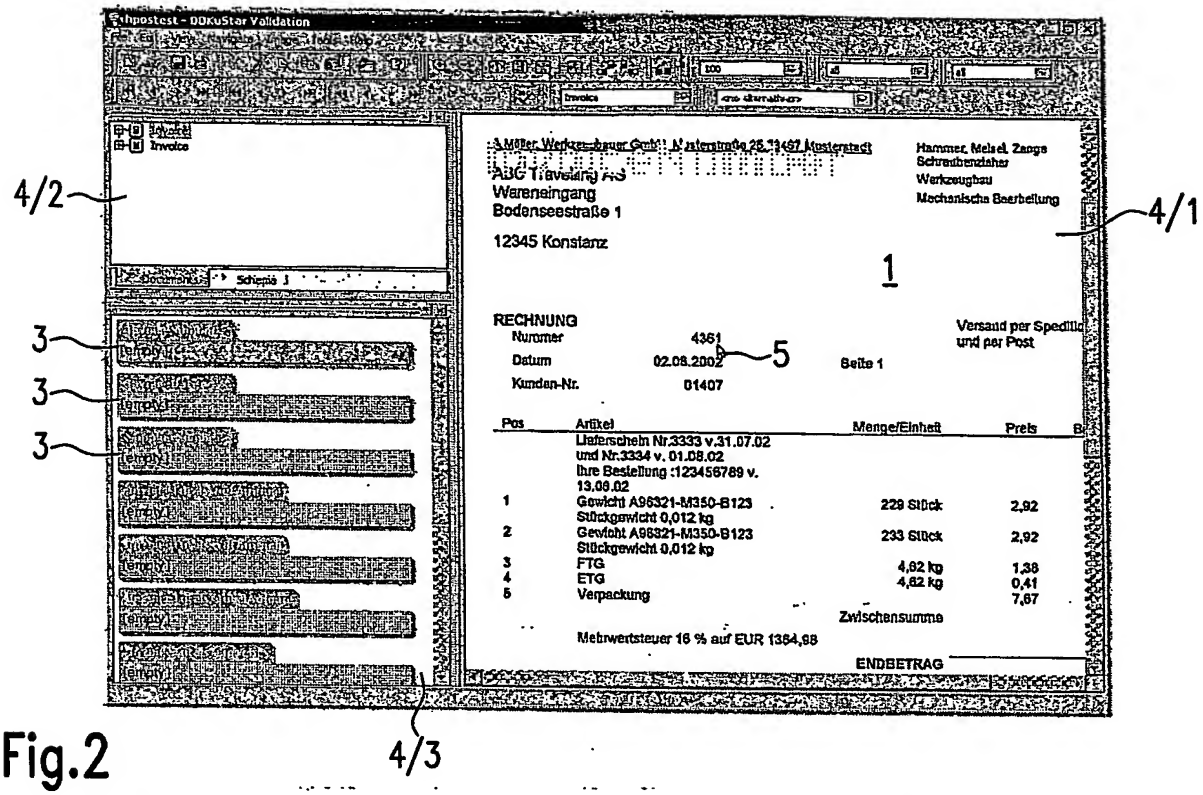


Fig.2

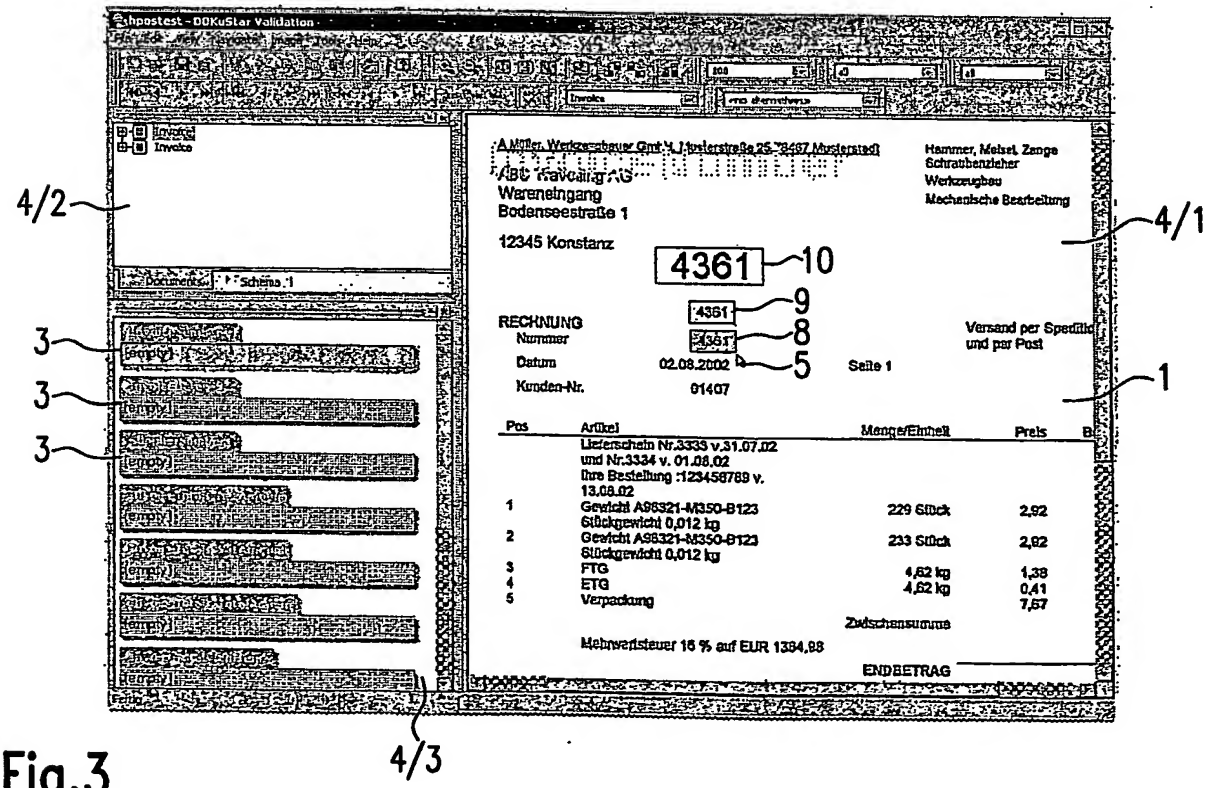


Fig.3

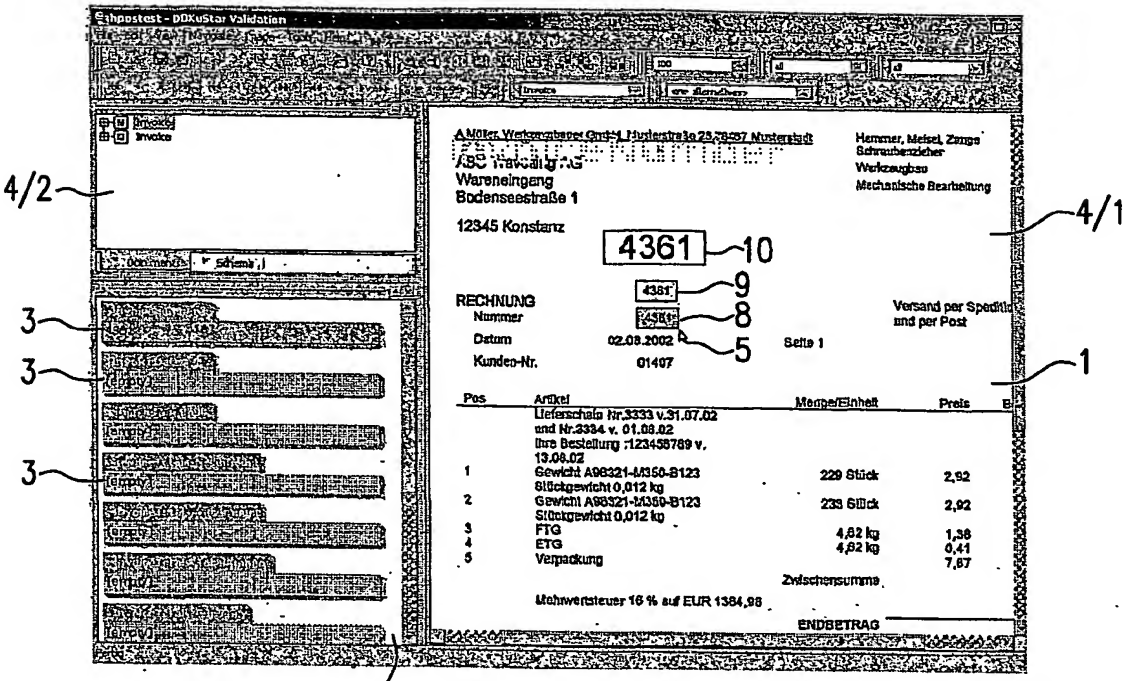

4/2
3
3
3
4/1
1
4/3
Warenеingang
Bodenseestraße 1
12345 Konstanz
4361
10
9
8
5
RECHNUNG
Nummer
Datum
02.08.2002
Seite 1
Kunden-Nr.
01407
Pos
Artikel
Menge/Einheit
Preis
Zwischensumme
Mehrwertsteuer 16 % auf EUR 1384,98
ENDBETRAG

Fig.4

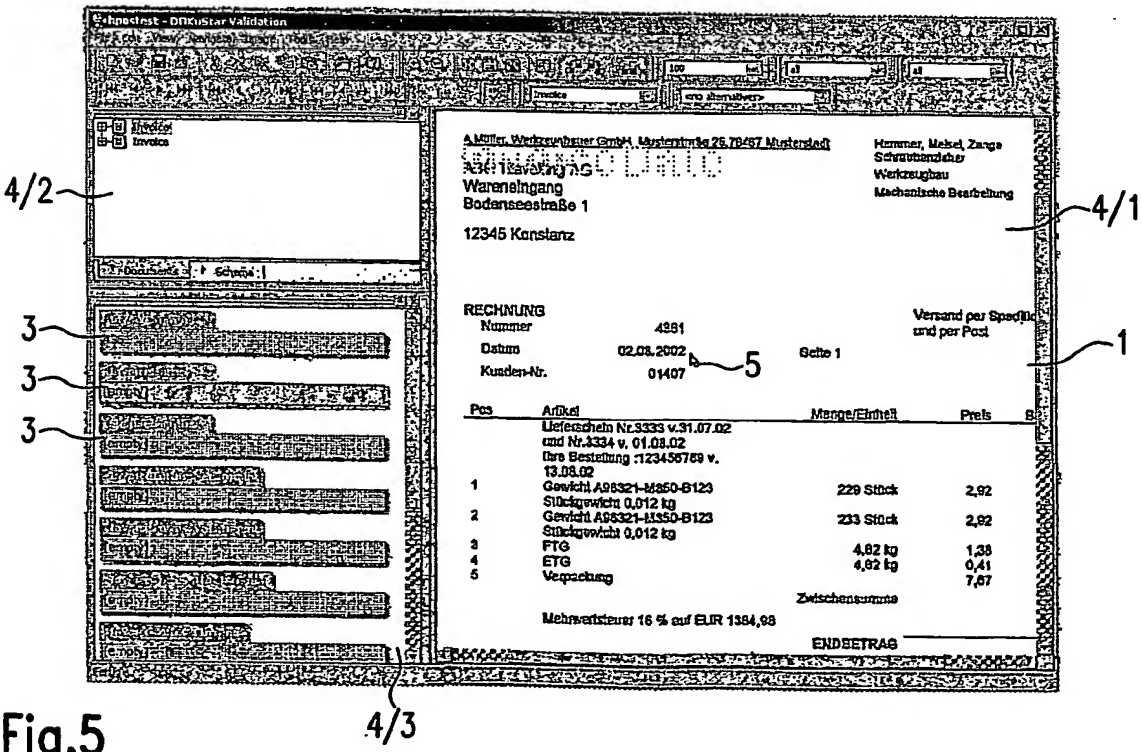

4/2
3
3
3
4/1
1
4/3
Warenеingang
Bodenseestraße 1
12345 Konstanz
RECHNUNG
Nummer
4361
Datum
02.08.2002
Seite 1
5
Kunden-Nr.
01407
Pos
Artikel
Menge/Einheit
Preis
Zwischensumme
Mehrwertsteuer 16 % auf EUR 1384,98
ENDBETRAG

Fig.5

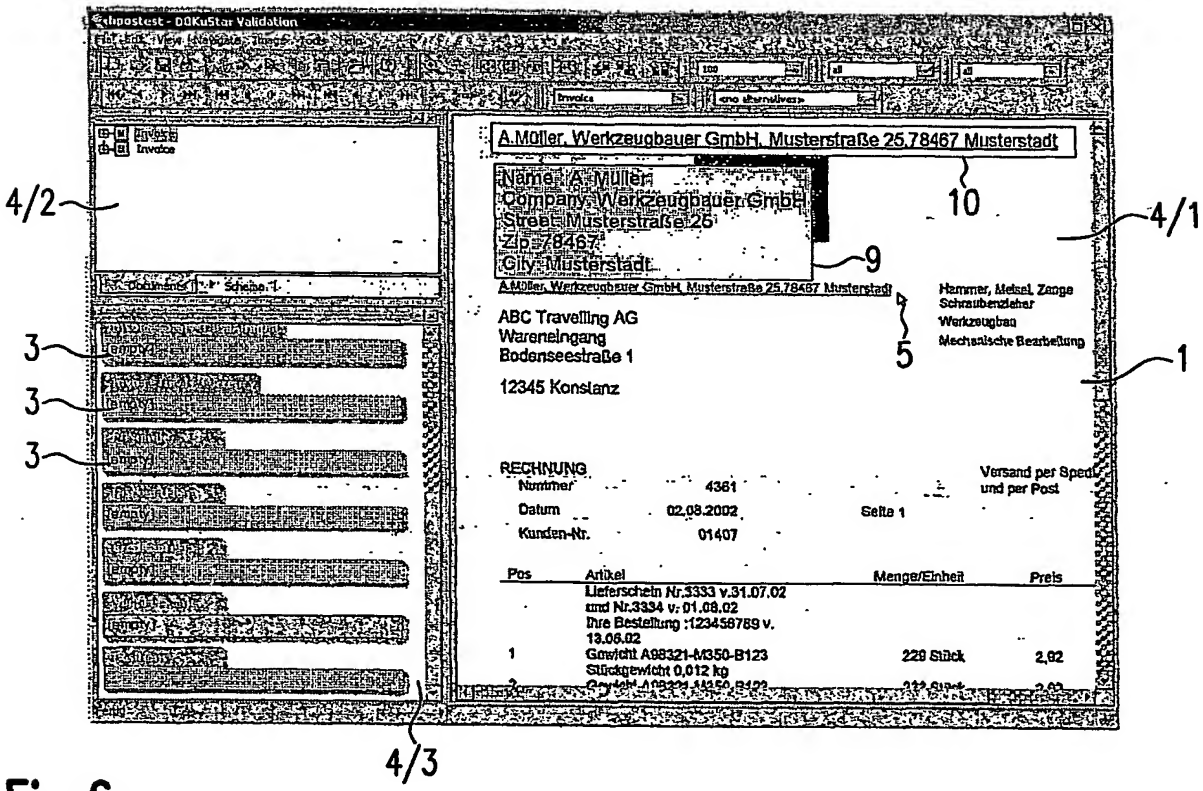

4/2
3
3
3
4/3
A.Müller, Werkzeugbauer GmbH, Musterstraße 25,78467 Musterstadt
10
4/1
Name A.Müller
Company Werkzeugbauer GmbH
Street Musterstraße 25
Zip 78467
City Musterstadt
9
A.Müller, Werkzeugbauer GmbH, Musterstraße 25,78467 Musterstadt
5
Hammer, Meisel, Zange
Schraubenzieher
Werkzeugbau
Mechanische Bearbeitung
1
ABC Travelling AG
Wareneingang
Bodenseestraße 1
12345 Konstanz
RECHNUNG
Nummer 4361
Datum 02.08.2002
Seite 1
Kunden-Nr. 01407
Versand per Spedi
und per Post
Pos
Artikel
Menge/Einheit
Preis
Lieferschein Nr.3333 v.31.07.02
und Nr.3334 v. 01.08.02
Ihre Bestellung :123456789 v.
13.06.02
1
Gewicht A98321-M350-B123
Stückgewicht 0,012 kg
229 Stück
2,62

Fig.6

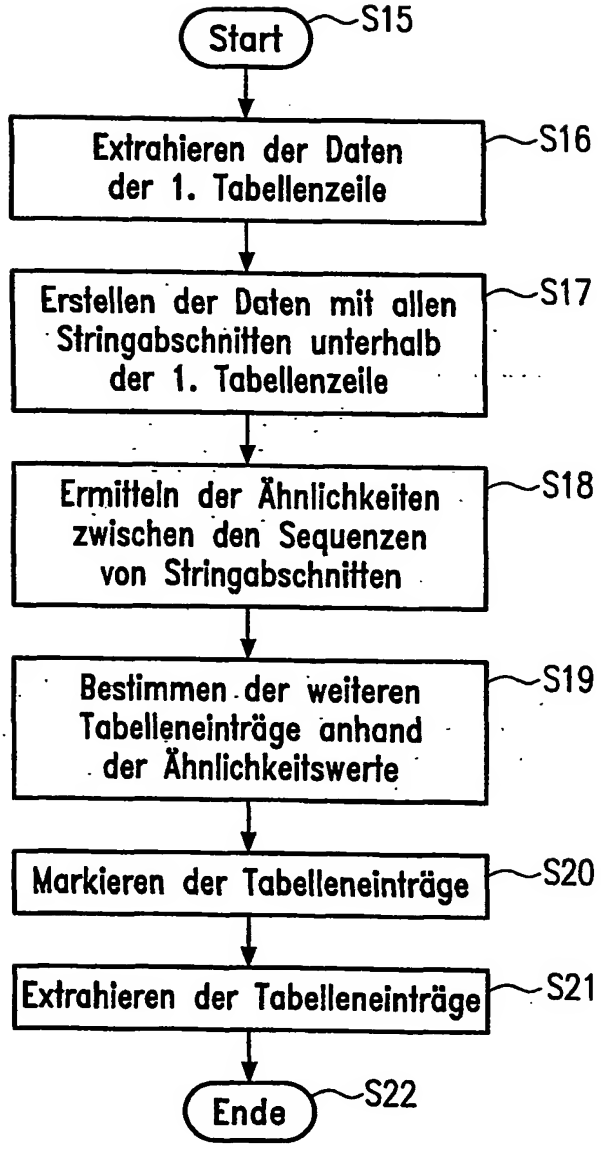
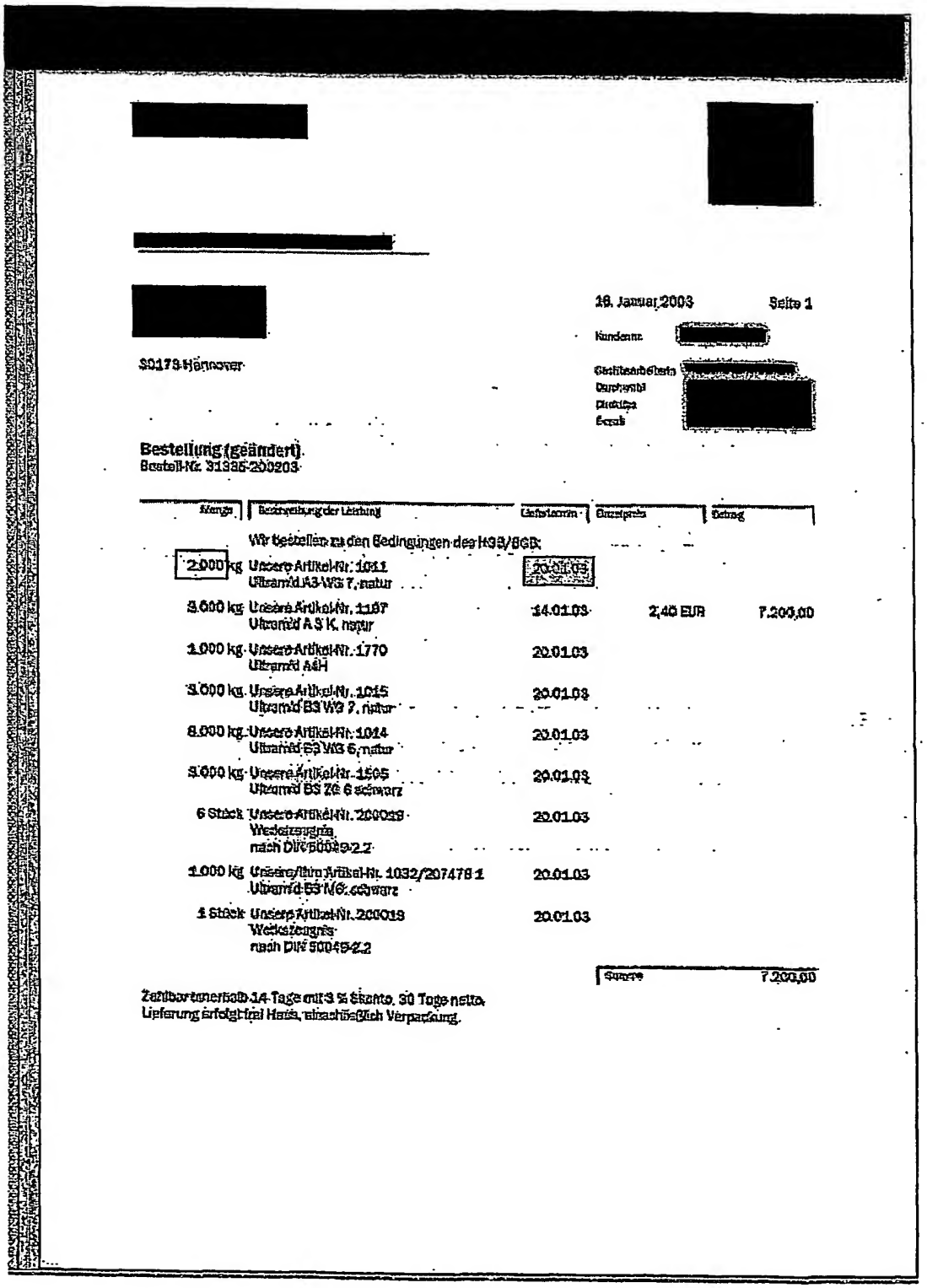

Start
S15
Extrahieren der Daten der 1. Tabellenzeile
S16
Erstellen der Daten mit allen Stringabschnitten unterhalb der 1. Tabellenzeile
S17
Ermitteln der Ähnlichkeiten zwischen den Sequenzen von Stringabschnitten
S18
Bestimmen der weiteren Tabelleneinträge anhand der Ähnlichkeitswerte
S19
Markieren der Tabelleneinträge
S20
Extrahieren der Tabelleneinträge
S21
Ende
S22

Fig.7



[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

30173 Hannover

19. Januar 2003 Seite 1

Kundennr. [redacted]

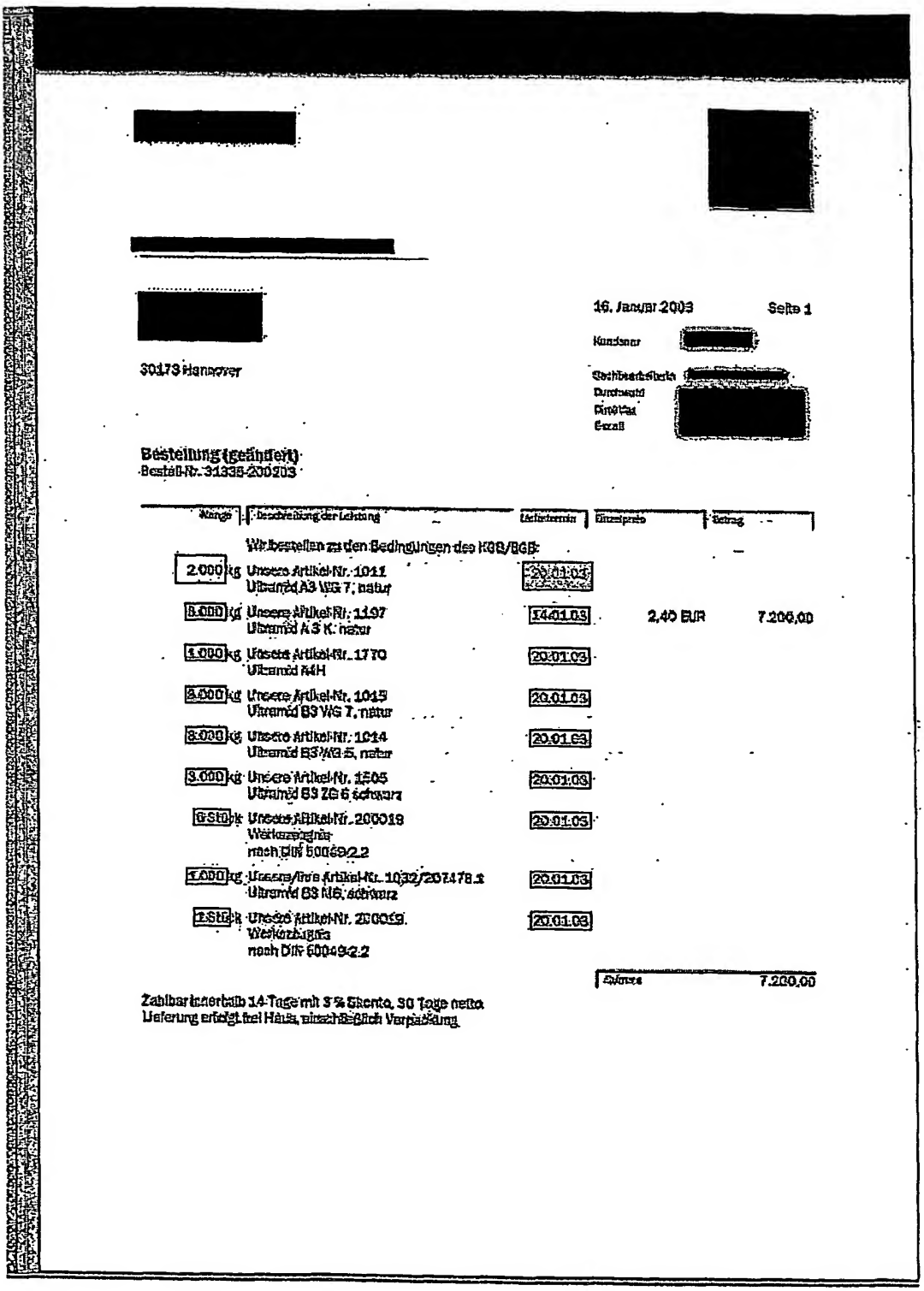
[redacted]

Bestellung (geändert)
Bestell-Nr. 31335-200203

Menge	Bezeichnung der Lieferung	Liefertermin	Einzelpreis	Betrag
	Wir bestellen zu den Bedingungen des HGB:			
2.000 kg	Unsere Artikel-Nr. 1011 Ultramid A3 W3 7, natur	20.01.03		
3.000 kg	Unsere Artikel-Nr. 1187 Ultramid A S K, natur	14.01.03	2,40 EUR	7.200,00
1.000 kg	Unsere Artikel-Nr. 1770 Ultramid A4H	20.01.03		
3.000 kg	Unsere Artikel-Nr. 1015 Ultramid B3 W9 7, natur	20.01.03		
8.000 kg	Unsere Artikel-Nr. 1014 Ultramid B3 W3 6, natur	20.01.03		
3.000 kg	Unsere Artikel-Nr. 1505 Ultramid B3 ZG 6 schwarz	20.01.03		
6 Stück	Unsere Artikel-Nr. 200019 Werkzeugnis nach DIN 50049-2.2	20.01.03		
1.000 kg	Unsere/Ihre Artikel-Nr. 1032/207478 1 Ultramid B3 W6 schwarz	20.01.03		
1 Stück	Unsere Artikel-Nr. 200019 Werkszeugnis nach DIN 50049-2.2	20.01.03		
			Summe	7.200,00

Zahlbar innerhalb 14 Tage mit 3 % Skonto, 30 Tage netto.
Lieferung erfolgt frei Haus, einschließlich Verpackung.

Fig.8



[redacted]

[redacted]

[redacted]

[redacted]

30173 Hannover

16. Januar 2003 Seite 1

Kundennr [redacted]

Sachbearbeiter/in [redacted]
Durchwahl
Ihr Fax
E-mail

[redacted]

Bestellung (geändert)
Bestell-Nr. 31335-200203

Menge	Bezeichnung der Lieferung	Lieferdatum	Einzelpreis	Betrag
	Wir bestellen zu den Bedingungen des HGB/BGB:			
2.000 kg	Unsere Artikel-Nr. 1011 Ultramid A3 WG 7, natur	20.01.03		
3.000 kg	Unsere Artikel-Nr. 1197 Ultramid A 3 K, natur	14.01.03	2,40 EUR	7.200,00
1.000 kg	Unsere Artikel-Nr. 1770 Ultramid A4H	20.01.03		
3.000 kg	Unsere Artikel-Nr. 1015 Ultramid B3 WG 7, natur	20.01.03		
3.000 kg	Unsere Artikel-Nr. 1014 Ultramid B3 WG 5, natur	20.01.03		
3.000 kg	Unsere Artikel-Nr. 1505 Ultramid B3 ZG 6 schwarz	20.01.03		
6 Stck	Unsere Artikel-Nr. 200019 Werkzeugbau nach DIN 50049-2.2	20.01.03		
1.000 kg	Unsere/Ihre Artikel-Nr. 1032/2074/8.1 Ultramid B3 NS, schwarz	20.01.03		
2 Stck	Unsere Artikel-Nr. 200019 Werkzeugbau nach DIN 50049-2.2	20.01.03		
			Summe	7.200,00

Zahlbar innerhalb 14 Tage mit 3 % Skonto, 30 Tage netto.
Lieferung erfolgt frei Haus, einschließlich Verpackung.

Fig.9

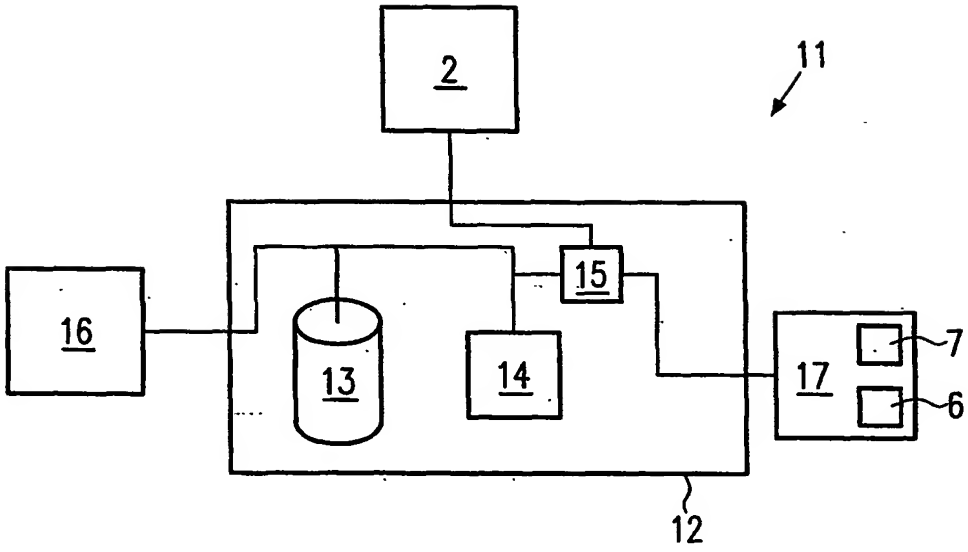


Fig.10

INTERNATIONAL SEARCH REPORT

International Application No
PCT/EP2004/009539

A. CLASSIFICATION OF SUBJECT MATTER
IPC 7 G06K9/20

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 7 G06K

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)
EPO-Internal

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 5 317 646 A (SANG JR HENRY W ET AL) 31 May 1994 (1994-05-31) Zusammenfassung Abschnitte: "Background of the invention", "Summary of the Invention" figures 1,2,4	1-19
X	US 2002/141660 A1 (PUCCI JORGE PABLO ET AL) 3 October 2002 (2002-10-03) Zusammenfassung Abschnitte: "Objects of the Invention", "Summary of the Invention"; Absätze: 0048-0051 claims 1,2; figures 1,2	1-19

-/--

[X] Further documents are listed in the continuation of box C.

[X] Patent family members are listed in annex.

° Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
16 February 2005	31/03/2005

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL – 2280 HV Rijswijk Tel. (+31–70) 340–2040, Tx. 31 651 epo nl, Fax (+31–70) 340–3016	Neubüser, B

INTERNATIONAL SEARCH REPORT

International Application No
PCT/EP2004/009539

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 6 028 970 A (DIPIAZZA PHILIP SILVANO ET AL) 22 February 2000 (2000-02-22) Zusammenfassung; column 7, line 47 - column 9, line 16; claims 1-5,14-19,27-29; figures 1A-C,2A,B,5	1-19
A	US 5 666 549 A (TSUCHIYA ET AL) 9 September 1997 (1997-09-09) claim 1; figures 2,5,6,16-18	12
A	US 5 966 473 A (ARIMA TOSHIMICHI ET AL) 12 October 1999 (1999-10-12) Abschnitte: "Background Art" "Summary of the Invention" figure 10	1-4,6, 15-19
A	CASEY R G ET AL: "INTELLIGENT FORMS PROCESSING" IBM SYSTEMS JOURNAL, IBM CORP. ARMONK, NEW YORK, US, vol. 29, no. 3, January 1990 (1990-01), pages 435-450, XP000265375 ISSN: 0018-8670 the whole document	1-19
A	WO 98/47098 A (ANDERSSON JAN ; READSOFT AB (SE)) 22 October 1998 (1998-10-22) cited in the application page 4, lines 5-7; figures 1,2	1

INTERNATIONAL SEARCH REPORT

Information on patent family members

International Application No
PCT/EP2004/009539

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 5317646 A	31-05-1994	NONE	
US 2002141660 A1	03-10-2002	NONE	
US 6028970 A	22-02-2000	NONE	
US 5666549 A	09-09-1997	JP 3002594 B2	24-01-2000
		JP 5290269 A	05-11-1993
		DE 4307577 A1	23-09-1993
		KR 123031 B1	21-11-1997
US 5966473 A	12-10-1999	JP 3113827 B2	04-12-2000
		JP 10162099 A	19-06-1998
WO 9847098 A	22-10-1998	SE 511242 C2	30-08-1999
		AT 247306 T	15-08-2003
		AU 6861798 A	11-11-1998
		DE 69817171 D1	18-09-2003
		DE 69817171 T2	17-06-2004
		DK 976092 T3	08-12-2003
		EP 0976092 A1	02-02-2000
		ES 2207824 T3	01-06-2004
		PT 976092 T	31-12-2003
		SE 9701183 A	02-10-1998
		WO 9847098 A1	22-10-1998

INTERNATIONALER RECHERCHENBERICHT

Internationales Aktenzeichen
PCT/EP2004/009539

A. KLASSIFIZIERUNG DES ANMELDUNGSGEGENSTANDES
IPK 7 G06K9/20

Nach der internationalen Patentklassifikation (IPK) oder nach der nationalen Klassifikation und der IPK

B. RECHERCHIERTE GEBIETE

Recherchierter Mindestprüfstoff (Klassifikationssystem und Klassifikationssymbole)
IPK 7 G06K

Recherchierte aber nicht zum Mindestprüfstoff gehörende Veröffentlichungen, soweit diese unter die recherchierten Gebiete fallen

Während der internationalen Recherche konsultierte elektronische Datenbank (Name der Datenbank und evtl. verwendete Suchbegriffe)
EPO-Internal

C. ALS WESENTLICH ANGESEHENE UNTERLAGEN

Kategorie°	Bezeichnung der Veröffentlichung, soweit erforderlich unter Angabe der in Betracht kommenden Teile	Betr. Anspruch Nr.
X	US 5 317 646 A (SANG JR HENRY W ET AL) 31. Mai 1994 (1994-05-31) Zusammenfassung Abschnitte: "Background of the invention", "Summary of the Invention" Abbildungen 1,2,4	1-19
X	US 2002/141660 A1 (PUCCI JORGE PABLO ET AL) 3. Oktober 2002 (2002-10-03) Zusammenfassung Abschnitte: "Objects of the Invention", "Summary of the Invention"; Absätze: 0048-0051 Ansprüche 1,2; Abbildungen 1,2	1-19

-/--

[X] Weitere Veröffentlichungen sind der Fortsetzung von Feld C zu entnehmen

[X] Siehe Anhang Patentfamilie

° Besondere Kategorien von angegebenen Veröffentlichungen :
"A" Veröffentlichung, die den allgemeinen Stand der Technik definiert, aber nicht als besonders bedeutsam anzusehen ist
"E" älteres Dokument, das jedoch erst am oder nach dem internationalen Anmeldedatum veröffentlicht worden ist
"L" Veröffentlichung, die geeignet ist, einen Prioritätsanspruch zweifelhaft erscheinen zu lassen, oder durch die das Veröffentlichungsdatum einer anderen im Recherchenbericht genannten Veröffentlichung belegt werden soll oder die aus einem anderen besonderen Grund angegeben ist (wie ausgeführt)
"O" Veröffentlichung, die sich auf eine mündliche Offenbarung, eine Benutzung, eine Ausstellung oder andere Maßnahmen bezieht
"P" Veröffentlichung, die vor dem internationalen Anmeldedatum, aber nach dem beanspruchten Prioritätsdatum veröffentlicht worden ist
"T" Spätere Veröffentlichung, die nach dem internationalen Anmeldedatum oder dem Prioritätsdatum veröffentlicht worden ist und mit der Anmeldung nicht kollidiert, sondern nur zum Verständnis des der Erfindung zugrundeliegenden Prinzips oder der ihr zugrundeliegenden Theorie angegeben ist
"X" Veröffentlichung von besonderer Bedeutung; die beanspruchte Erfindung kann allein aufgrund dieser Veröffentlichung nicht als neu oder auf erfinderischer Tätigkeit beruhend betrachtet werden
"Y" Veröffentlichung von besonderer Bedeutung; die beanspruchte Erfindung kann nicht als auf erfinderischer Tätigkeit beruhend betrachtet werden, wenn die Veröffentlichung mit einer oder mehreren anderen Veröffentlichungen dieser Kategorie in Verbindung gebracht wird und diese Verbindung für einen Fachmann naheliegend ist
"&" Veröffentlichung, die Mitglied derselben Patentfamilie ist

Datum des Abschlusses der internationalen Recherche	Absendedatum des internationalen Recherchenberichts
16. Februar 2005	31/03/2005

Name und Postanschrift der Internationalen Recherchenbehörde	Bevollmächtigter Bediensteter
Europäisches Patentamt, P.B. 5818 Patentlaan 2 NL – 2280 HV Rijswijk Tel. (+31–70) 340–2040, Tx. 31 651 epo nl, Fax: (+31–70) 340–3016	Neubüser, B

INTERNATIONALER RECHERCHENBERICHT

Internationales Aktenzeichen
PCT/EP2004/009539

C.(Fortsetzung) ALS WESENTLICH ANGESEHENE UNTERLAGEN

Kategorie°	Bezeichnung der Veröffentlichung, soweit erforderlich unter Angabe der in Betracht kommenden Teile	Betr. Anspruch Nr.
X	US 6 028 970 A (DIPIAZZA PHILIP SILVANO ET AL) 22. Februar 2000 (2000-02-22) Zusammenfassung; Spalte 7, Zeile 47 - Spalte 9, Zeile 16; Ansprüche 1-5,14-19,27-29; Abbildungen 1A-C,2A,B,5	1-19
A	US 5 666 549 A (TSUCHIYA ET AL) 9. September 1997 (1997-09-09) Anspruch 1; Abbildungen 2,5,6,16-18	12
A	US 5 966 473 A (ARIMA TOSHIMICHI ET AL) 12. Oktober 1999 (1999-10-12) Abschnitte: "Background Art" "Summary of the Invention" Abbildung 10	1-4,6, 15-19
A	CASEY R G ET AL: "INTELLIGENT FORMS PROCESSING" IBM SYSTEMS JOURNAL, IBM CORP. ARMONK, NEW YORK, US, Bd. 29, Nr. 3, Januar 1990 (1990-01), Seiten 435-450, XP000265375 ISSN: 0018-8670 das ganze Dokument	1-19
A	WO 98/47098 A (ANDERSSON JAN ; READSOFT AB (SE)) 22. Oktober 1998 (1998-10-22) in der Anmeldung erwähnt Seite 4, Zeilen 5-7; Abbildungen 1,2	1

INTERNATIONALER RECHERCHENBERICHT

Angaben zu Veröffentlichungen, die zur selben Patentfamilie gehören

Internationales Aktenzeichen
PCT/EP2004/009539

Im Recherchenbericht angeführtes Patentdokument		Datum der Veröffentlichung	Mitglied(er) der Patentfamilie		Datum der Veröffentlichung
US 5317646	A	31-05-1994	KEINE		
US 2002141660	A1	03-10-2002	KEINE		
US 6028970	A	22-02-2000	KEINE		
US 5666549	A	09-09-1997	JP	3002594 B2	24-01-2000
			JP	5290269 A	05-11-1993
			DE	4307577 A1	23-09-1993
			KR	123031 B1	21-11-1997
US 5966473	A	12-10-1999	JP	3113827 B2	04-12-2000
			JP	10162099 A	19-06-1998
WO 9847098	A	22-10-1998	SE	511242 C2	30-08-1999
			AT	247306 T	15-08-2003
			AU	6861798 A	11-11-1998
			DE	69817171 D1	18-09-2003
			DE	69817171 T2	17-06-2004
			DK	976092 T3	08-12-2003
			EP	0976092 A1	02-02-2000
			ES	2207824 T3	01-06-2004
			PT	976092 T	31-12-2003
			SE	9701183 A	02-10-1998
			WO	9847098 A1	22-10-1998